

IDENTIFIABILITY OF LARGE PHYLOGENETIC MIXTURE MODELS

JOHN A. RHODES AND SETH SULLIVANT

ABSTRACT. Phylogenetic mixture models are statistical models of character evolution allowing for heterogeneity. Each of the classes in some unknown partition of the characters may evolve by different processes, or even along different trees. The fundamental question of whether parameters of such a model are identifiable is difficult to address, due to the complexity of the parameterization. We analyze mixture models on large trees, with many mixture components, showing that both numerical and tree parameters are indeed identifiable in these models when all trees are the same. We also explore the extent to which our algebraic techniques can be employed to extend the result to mixtures on different trees.

1. INTRODUCTION

A fundamental question about any parametric statistical model is whether or not the parameters of that model are *identifiable*; that is, does a probability distribution arising from the model uniquely determine the parameters that produced it. Establishing the identifiability of parameters is important for statistical inference, especially in models where the parameters have a physical or biological interpretation. For example, it is well-known that identifiability is a necessary condition for statistical consistency of maximum likelihood estimation [14, Chapter 16].

In phylogenetics, parameters of interest include the discrete tree parameter and numerical parameters specifying substitution processes on the edges of the tree. For the simplest phylogenetic models, identifiability of both tree and numerical parameters have long been established [11]. But as models grow in complexity, with both the combinatorial description of trees and the underlying number of numerical parameters increasing, the question of identifiability is far from settled.

A particular class of complex phylogenetic models of growing interest and use are the phylogenetic mixture models. Relatively simple examples are the models with small numbers of parameters — including those with Γ -distributed rates, invariable sites, and combinations of these — that are currently the most commonly used in data analysis. More elaborate mixtures allow across-site rate variation with more freedom in the distribution of the rate multipliers [15], the use of different rate matrices [18], or even multiple distinct trees each with their own rate and time parameters. Such models may have a large number of mixture components. For instance, a Bayesian nonparametric analysis conducted in [15] allowed a variable number of components, with a Dirichlet process prior specifying a mean of as many as 20.

However, only the simplest phylogenetic mixture models have been proven to be identifiable, typically where the number of parameters are small. The papers [1, 4, 6, 7, 10, 21]

contain previous results on identifiability of such models, of various sorts. Note that only recently has it been shown that most choices of parameters of the widely-used GTR + I + Γ model are identifiable [10], although for a certain type of rate matrix the question remains open.

Our goal in this paper is to develop methods to prove identifiability in phylogenetic models that are considerably more complex than in previous work. In particular, we investigate the identifiability of phylogenetic models with many mixing components. A consequence of our methods is the following theorem:

Theorem 1.1. *For an r -component identical tree mixture of the general Markov model of character evolution with κ -state random variables on an n -leaf binary phylogenetic tree, both the tree parameter and the numerical parameters are generically identifiable if $r < \kappa^{\lceil n/4 \rceil - 1}$.*

By an *identical tree mixture model* we mean a mixture of probability distributions coming from the same topological phylogenetic tree. More complicated mixture models might have each distribution arising from a different topological tree. Theorem 1.1 improves substantially over past identifiability results on identical tree phylogenetic mixture models. Previously, it was only known that the tree parameter is identifiable, and then only in the case that $r < \kappa$ (work of Allman and the first author [6]). This new theorem quantifies the intuition that larger taxon sets should allow for identifiability of more complex models, and is an exponential improvement over previous results.

Our strategy of proof is to combine two techniques coming from the algebraic study of phylogenetic models. First, we use the representation of probability distributions in a phylogenetic model as tensors with small tensor rank and employ a theorem of J. Kruskal to uniquely identify components of that tensor. Second, we use phylogenetic invariants as tools to identify deeply embedded features of phylogenetic trees, and to “untangle” probability distributions that have been shuffled together by the tensor analysis. While each technique by itself is only able to make a small advance on the identifiability problem, when combined they give dramatically stronger results. Background on these general techniques appears in Section 3, and the proofs of the main theorems are in Section 4.

Our techniques actually extend to mixtures from different trees provided they all share a certain type of common substructure. It is in this generality that we prove our main results, Theorems 4.6 and 4.7, with Theorem 1.1 arising as a corollary.

The assumption of any common substructure in the trees is of course false in some biological situations modeled by mixtures. For instance, if the mixture is due to the coalescent process modeling incomplete lineage sorting on a species tree of populations, then components will be present from all topological gene trees [13, 22]. However, one might also model lateral gene transfer at a number of (unknown) locations in a tree as a mixture, and for this the assumption of common substructure could be quite plausible.

2. PRELIMINARIES

2.1. Mixture models. Consider the general Markov model of κ -state character evolution, $GM(\kappa)$, on n -taxon trees (*e.g.*, $\kappa = 4$ corresponding to DNA sequences). We assume

the taxa labeling the leaves are identified with $[n] = \{1, 2, \dots, n\}$. Then for each rooted leaf-labeled tree T , there is a *parametrization map* ψ_T giving the joint distribution of states at the leaves of the tree T as functions of continuous parameters, which specify the state distribution at the root and the transition probabilities on the edges. Let S_T denote the continuous parameter space of $GM(\kappa)$ on T , which is a full dimensional subset of some \mathbb{R}^m . Then

$$\psi_T : S_T \rightarrow \Delta^{\kappa^n - 1},$$

where $\Delta^{\ell-1} \subseteq [0, 1]^\ell$ denotes the probability simplex comprised of non-negative real vectors summing to 1. The image of this map is the phylogenetic model $\mathcal{M}_T \subseteq \Delta^{\kappa^n - 1}$.

The associated r -component mixture model has the following parametrization: For every r -tuple of trees $\mathbf{T} = (T_1, T_2, \dots, T_r)$ on the same taxa $[n]$, let $S_{\mathbf{T}} = S_{T_1} \times \dots \times S_{T_r} \times \Delta^{r-1}$ and let

$$\psi_{\mathbf{T}} : S_{\mathbf{T}} \rightarrow \Delta^{\kappa^n - 1},$$

be defined by

$$\psi_{\mathbf{T}}(s_1, \dots, s_r, \pi) = \pi_1 \psi_{T_1}(s_1) + \dots + \pi_r \psi_{T_r}(s_r).$$

Thus π is the vector of mixing parameters; each π_i gives the proportion of i.i.d. sites that evolve along tree T_i with parameter vector s_i . The r -component mixture model on \mathbf{T} is the image of the map $\psi_{\mathbf{T}}$, and is denoted

$$\mathcal{M}_{\mathbf{T}} = \mathcal{M}_{T_1} * \mathcal{M}_{T_2} * \dots * \mathcal{M}_{T_r}.$$

Clearly $\mathcal{M}_{\mathbf{T}}$ depends only on the unordered multiset of the trees in \mathbf{T} . In the case where $T_i = T$ for all i , we call this an r -component *identical tree* mixture model on T .

We focus on the mixture models built from the basic model $GM(\kappa)$ in this paper, as these are quite general *algebraic models*, for which the maps ψ_T are naturally defined by polynomial formulas. Many models which are not polynomial (in particular, those built from the general time-reversible model) can be embedded in them. The polynomial structure of algebraic models allows them to be studied using techniques from algebraic geometry.

2.2. Identifiability of parameters. For algebraic models, it is convenient to slightly weaken the notion of identifiability of parameters to *generic identifiability*. The word “generic” is used to mean “except on a proper algebraic subvariety” of the parameter space. (See section 3.3 for a formal definition of variety.) Although it is sometimes possible to be explicit about this subvariety, we usually are not, since the key point in interpretation is that a proper subvariety is a closed set of Lebesgue measure 0 inside the larger set. Thus regardless of the precise subvariety involved, randomly chosen points are generic with probability 1.

On an unmixed $GM(\kappa)$ model on a single tree T , there are several well-understood issues with identifiability of parameters. First, at any internal node of the tree, in a phenomenon called *label swapping*, one may permute the names of the state space of the corresponding hidden variable (permuting the columns or rows of the Markov matrices on edges leading to or from the node) with no effect on the probability distribution. Second, while the standard parameterization of the $GM(\kappa)$ model on a tree T requires specification of the

root of T , for generic choices of parameters one can relocate the root (with an appropriate uniquely determined change to the parameters, up to label swapping) with no effect on the probability distribution. Third, if any internal nodes of T have degree 2, they may be suppressed and the Markov matrices on incident edges combined, with no effect on the probability distribution. Thus one generally assumes trees have no such nodes. For simplicity, we do not always explicitly refer to these issues in our formal statements in this article. However, we will occasionally use the second fact to choose a convenient location for a root of a tree in our arguments.

That these are the only issues for parameter identifiability for the unmixed model is the content of the following theorem, which was essentially shown in [11].

Theorem 2.1. *For the $GM(\kappa)$ model on a single tree,*

- (1) *The unrooted tree parameter is generically identifiable, in the class of binary trees.*
- (2) *For a fixed binary tree T , the numerical parameters of the $GM(\kappa)$ model on T are generically identifiable, up to label swapping at internal nodes of the tree, and an arbitrary choice of a node as the root.*

An additional issue for identifiability of r -tree mixtures is component swapping: Interchanging the trees along with their parameters, while permuting the mixing parameters in the same way, has no effect on the resulting distribution. A useful notion of identifiability must allow for this.

Definition 2.2. The tree parameters of the r -tree mixture are generically identifiable if for any binary trees $\mathbf{T} = (T_1, \dots, T_r)$ on the same set of taxa, and generic choices of parameters s_1, \dots, s_r, π ,

$$\psi_{\mathbf{T}}(s_1, \dots, s_r, \pi) = \psi_{\mathbf{T}'}(s'_1, \dots, s'_r, \pi')$$

implies that $\mathbf{T} = \sigma \cdot \mathbf{T}'$ for some $\sigma \in \mathfrak{S}_r$, the symmetric group of permutations.

We also investigate identifiability of tree parameters when restricting to specific classes of r -tuples of trees. For example, Theorem 1.1 concerns identifiability of tree parameters among all sets $\mathbf{T} = \{T_1, \dots, T_r\}$, where $T_1 = \dots = T_r$. Our main results, Theorems 4.6 and 4.7 concern identifiability in the class of r -tuples of trees that all contain a specified deep common substructure, whose precise definition will be given in Section 4.

Definition 2.3. The continuous parameters of an r -tree mixture on \mathbf{T} are generically identifiable if for generic choices of s_1, \dots, s_r and π ,

$$\psi_{\mathbf{T}}(s_1, \dots, s_r, \pi) = \psi_{\mathbf{T}'}(s'_1, \dots, s'_r, \pi')$$

implies that there is a permutation $\sigma \in \mathfrak{S}_r$ such that $\sigma \cdot \mathbf{T} = \mathbf{T}$, $s'_i = s_{\sigma(i)}$, and $\pi'_i = \pi_{\sigma(i)}$ for $i = 1, \dots, r$.

Note this definition only allows the swapping of continuous parameters s_i, π_i with s_j, π_j when $T_i = T_j$.

2.3. Splits and tripartitions. We will use the combinatorial notion of a split of the leaves of a tree associated to an edge in a binary tree, as well as the analog of this concept for a node of the tree.

Definition 2.4. A split of $[n]$ is a bipartition $A|B$ of $[n]$ with two nonempty elements. A split is said to be compatible with a tree T if it arises as the partition of leaves induced by an edge in some binary resolution of T .

Similarly, a tripartition of $A|B|C$ of leaves is said to be compatible with T if it arises as the tripartition induced by an interior vertex in some binary resolution of T .

A collection of trees is said to have a common split (or tripartition) if the split (or tripartition) is compatible with every tree in the collection.

A collection of trees has a common tripartition $A|B|C$ if, and only if, it also has the three common splits $A|B \cup C$, $B|A \cup C$, and $C|A \cup B$. For a binary tree, these are the splits associated to the edges radiating from the vertex inducing the tripartition. Note also that our definition of compatible splits differs from the standard definition (*e.g.*, in [19]) in the case of trees with polytomies. Our notion is more useful when studying geometric properties of phylogenetic models.

3. TENSORS AND INVARIANTS

The two main tools we use to prove our results are Kruskal's theorem on uniqueness of tensor decompositions and phylogenetic invariants. In this section, we describe these tools. Both are connected to the notion of a flattening of the probability distribution arising from a phylogenetic model.

3.1. Tensors and Unique Decomposition. By a *tensor*, we mean simply an n -way rectangular array of numbers. A 2-way tensor is thus a matrix.

For $j = 1, 2, 3$, let M_j be an $r \times \kappa_j$ matrix with i th row $\mathbf{m}_i^j = (m_i^j(1), \dots, m_i^j(\kappa_j))$. Let $[M_1, M_2, M_3]$ denote the 3-way $\kappa_1 \times \kappa_2 \times \kappa_3$ tensor defined by

$$[M_1, M_2, M_3] = \sum_{i=1}^r \mathbf{m}_i^1 \otimes \mathbf{m}_i^2 \otimes \mathbf{m}_i^3.$$

In other words, $[M_1, M_2, M_3]$ is an $\kappa_1 \times \kappa_2 \times \kappa_3$ array whose (u, v, w) entry is

$$[M_1, M_2, M_3]_{u,v,w} = \sum_{i=1}^r m_i^1(u) m_i^2(v) m_i^3(w).$$

Every 3-way tensor can be expressed in this way, for sufficiently large r . A nonzero tensor of this form with $r = 1$ is said to have tensor rank 1. More generally, the minimal r such that a 3-way tensor can be decomposed as such a sum is called its *tensor rank*. A natural question is when this expression is essentially unique.

Note there are two basic operations on the matrices M_1, M_2, M_3 which leave unchanged the tensor $[M_1, M_2, M_3]$: one can simultaneously permute the rows of the three matrices M_1, M_2 , and M_3 , or taking three numbers a_1, a_2, a_3 such that $a_1 a_2 a_3 = 1$, one can replace the i th rows \mathbf{m}_i^j by $a_j \mathbf{m}_i^j$. Kruskal's Theorem [16, 17] describes a situation where these operations lead to the only variants in a tensor decomposition.

Given an $r \times \kappa$ matrix M , its *Kruskal rank*, denoted $\text{rank}_K(M)$, is the largest value k such that every subset of k rows of M is linearly independent. Note that $\text{rank}_K(M) \leq \text{rank}(M)$.

Theorem 3.1 ([16, 17]). *Let $I_j = \text{rank}_K(M_j)$, where M_j is $r \times \kappa_j$. If*

$$I_1 + I_2 + I_3 \geq 2r + 2$$

then $[M_1, M_2, M_3]$ uniquely determines M_1, M_2, M_3 up to simultaneous permutation and scaling of the rows.

Kruskal's theorem has proven useful for proving identifiability results of numerical parameters for both phylogenetic models [9] and for other statistical models with hidden variables [2, 3]. We will show how to combine this with other algebraic techniques to also deduce identifiability of tree parameters.

3.2. Flattenings. While Kruskal's theorem concerns 3-way tensors, the tensors arising in phylogenetics are usually n -way $\kappa \times \cdots \times \kappa$ tensors, corresponding to the n leaves of a phylogenetic tree. We will make frequent use of flattenings of n -way tensors to lower order tensors. A flattening of a n -way tensor is simply a reorganization of that tensor as a k -way tensor, with $k < n$, of larger dimensions. We take a $\kappa_1 \times \cdots \times \kappa_n$ tensor M , with typical entry $M(u_1, \dots, u_n)$, and a partition $A_1|A_2|\cdots|A_k$ of $[n]$, and we represent this as a

$$\prod_{a \in A_1} \kappa_a \times \cdots \times \prod_{a \in A_k} \kappa_a$$

tensor \tilde{M} . The (u_1, \dots, u_n) entry of M becomes the $((u_a)_{a \in A_1}, \dots, (u_a)_{a \in A_k})$ entry of \tilde{M} . That is, the indices for the new tensor \tilde{M} are vectors of indices from the tensor M .

Given a partition $A_1|A_2|\cdots|A_k$ of $[n]$, we denote the corresponding flattening of M by $\text{Flat}_{A_1|A_2|\cdots|A_k}(M)$.

3.3. Invariants, Phylogenetic and Otherwise. We begin with a little background on algebraic geometry (see [12] for more detail). Let $\mathbb{R}[p_1, \dots, p_m]$ be the set of all polynomials in the variables (or indeterminates) p_1, p_2, \dots, p_m , with coefficients in the real numbers, \mathbb{R} . Algebraic geometry studies the zero sets of collections of polynomials. That is, to a collection of polynomials $f_1, f_2, \dots, f_k \in \mathbb{R}[p_1, \dots, p_m]$ we associate the *variety*

$$V(f_1, \dots, f_k) = \{\mathbf{a} \in \mathbb{R}^m : f_1(\mathbf{a}) = f_2(\mathbf{a}) = \cdots = f_k(\mathbf{a}) = 0\}.$$

The fact that these geometric sets arise from polynomials vanishing implies they have important structural features.

Varieties arise in studying statistical models through describing models implicitly, rather than parametrically. For a fixed statistical model $\mathcal{M} \subseteq \Delta^{m-1}$, an *invariant* of \mathcal{M} is a polynomial $f \in \mathbb{R}[p_1, \dots, p_m]$ such that $f(\mathbf{a}) = 0$ for all $\mathbf{a} \in \mathcal{M}$. In the case where \mathcal{M} is a phylogenetic model, such a polynomial is called a *phylogenetic invariant*.

Our main use in this paper for phylogenetic invariants is their connection to generic identifiability, through the following basic proposition from algebraic geometry.

Proposition 3.2. *Let V_0 and V_1 be two irreducible algebraic varieties, such as those arising from parameterized statistical models. Suppose f_0 is an invariant for V_0 , and*

there exists a point $p_1 \in V_1$ with $f_0(p_1) \neq 0$. Then $V_1 \not\subseteq V_0$, and the variety $V_0 \cap V_1$ is of lower dimension than V_1 . That is, generic points on V_1 lie off of V_0 .

Among the most important and elementary phylogenetic invariants are the ones that arise from edge flattenings of tensors.

Definition 3.3. Let $A|B$ be a split compatible with the tree T . An *edge invariant* for T is a phylogenetic invariant that can be expressed as a minor (i.e., the determinant of a submatrix) of the matrix $\text{Flat}_{A|B}(P)$.

As an indication of how edge invariants can be used to identify combinatorial information on the tree underlying a phylogenetic model, we recall the following theorem concerning models on a single tree. While this statement is well-known in the phylogenetic invariants literature, Theorem 4.1 of this article provides a more general extension to mixture models.

Theorem 3.4. Suppose that T_0 and T_1 are two n -leaf trees such that for $i = 0, 1$, $A_i|B_i$ is a split compatible with T_i and incompatible with T_{1-i} , and let \mathcal{M}_i denote the κ -state general Markov model $GM(\kappa)$ on T_i . Then the $(\kappa + 1)$ -minors of $\text{Flat}_{A_i|B_i}(P)$ vanish on \mathcal{M}_{T_i} and do not vanish on $\mathcal{M}_{T_{1-i}}$, and thus are edge invariants for the first model but not the second. In particular, edge invariants can be used to generically identify the tree topology.

Edge invariants have been the phylogenetic invariants most interesting for tree identifiability in the past, and contain enough information to reconstruct the combinatorial type of a single tree in some situations. However, we need some more complicated invariants to get more information in the case of the phylogenetic mixture models considered here. We describe these invariants, discovered in several different contexts [5, 20], in matrix form.

Theorem 3.5. Let P be a $\kappa \times \kappa \times \kappa$ tensor giving a distribution from the $GM(\kappa)$ model on a 3-leaf tree. For $i = 1, \dots, \kappa$, let $P_{(i)}$ be the matrix slice $P_{(i)} = (P(i, u, v))_{u,v}$. Then

$$P_{(i)} (\text{adj } P_{(j)}) P_{(k)} - P_{(k)} (\text{adj } P_{(j)}) P_{(i)} = 0.$$

Here $\text{adj } A$ denotes the classical adjoint of A , which is given by polynomial expressions in the entries of A . In the case of nonsingular A , $\text{adj}(A) = \det(A)A^{-1}$.

4. IDENTIFIABILITY OF MIXTURE MODELS WITH COMMON SUBSTRUCTURE

In this section, we prove our main result, that both tree parameters and numerical parameters are generically identifiable in a phylogenetic mixture model provided we restrict to multisets \mathbf{T} of trees that all share a certain substructure. More precisely, we require that all trees in \mathbf{T} have two splits in common. The number of mixing components that can be identified via our techniques will depend on the sizes of the sets in these splits. As a corollary, we deduce Theorem 1.1, after showing that if all trees are the same, there is a “deep” internal vertex with two of its incident edges giving the requisite splits.

Before proceeding to the statements and proofs of the main theorems, we prove three lemmas.

Lemma 4.1. (*Edge invariants for tree mixtures*)

Consider the $GM(\kappa)$ mixture model on r trees $\mathbf{T} = (T_1, \dots, T_r)$. Let $A|B$ be a bipartition of the taxa, with $r < \min(\kappa^{\#A-1}, \kappa^{\#B-1})$

- (1) If $A|B$ is compatible with all trees in \mathbf{T} , then all $(r\kappa + 1)$ -minors of $\text{Flat}_{A|B}(P)$ vanish for all distributions P arising from the model.
- (2) If $A|B$ is not compatible with at least one tree in \mathbf{T} , then for generic distributions P arising from the model at least one $(r\kappa + 1)$ -minor of $\text{Flat}_{A|B}(P)$ does not vanish.

Proof. The claims concerning (non)vanishing of minors are equivalent to claims that $\text{Flat}_{A|B}(P)$ has rank at most $r\kappa$ in case (1), and generically has rank greater than $r\kappa$ in case (2). Therefore we focus on investigating ranks of flattenings.

If $A|B$ is compatible with all trees in \mathbf{T} , then, by passing to binary resolutions of the T_i , we may assume it is a split associated to edge $e_i = (a_i, b_i)$ in T_i . Then one sees that

$$\text{Flat}_{A|B}(P) = M_A^T Q M_B.$$

Here Q is the $r\kappa \times r\kappa$ block-diagonal matrix whose i th $\kappa \times \kappa$ block gives the joint probability distribution of states for the random variables at a_i and b_i , weighted by the component proportion π_i . The matrices M_A, M_B are stochastic, of sizes $r\kappa \times \kappa^{\#A}$, $r\kappa \times \kappa^{\#B}$, with entries in the i th block of κ rows giving probabilities of states of variables in A, B conditioned on states at a_i, b_i . This factorization implies the claimed bound on the rank.

Suppose next that $A|B$ is not compatible with at least one of the trees in \mathbf{T} , say T_1 . To show that $\text{Flat}_{A|B}(P)$ generically has rank greater than $r\kappa$, it is enough to give a single choice of parameters producing such a rank. Indeed, this follows from Proposition 3.2, applied to the model and the variety of matrices of rank at most $r\kappa$.

To simplify this choice, for each T_i with $i > 1$ choose all Markov matrices for all internal edges of T_i to be the identity, I_κ . Since T_1 is not compatible with $A|B$, by Theorem 3.8.6 of [19], it has an edge $e = (c, d)$, with associated split $C|D$, such that all four sets $A \cap C$, $A \cap D$, $B \cap C$, $B \cap D$ are nonempty. For all internal edges of T_1 except e , choose Markov matrices to be I_κ as well. Since the effect of an identity matrix on an edge is the same as contracting that edge, with these choices we need henceforth argue only in the following special case: for $i > 1$, T_i is a star tree with central node a_i , and T_1 has the form of two star trees, on C and on D , that are joined at their central nodes by e .

Now express the distribution $P = P_1 + P'$ where P_1 is the mixture component from T_1 , and P' the sum of the components on the star trees $T_2 = \dots = T_r$. Then, one can write

$$M_2 := \text{Flat}_{A|B}(P') = N_A^T R N_B,$$

with R an $(r-1)\kappa \times (r-1)\kappa$ diagonal matrix giving the distribution of states at a_i in components $2, \dots, r$ weighted by the π_i , and N_A, N_B are stochastic matrices of sizes $(r-1)\kappa \times \kappa^{\#A}$, $(r-1)\kappa \times \kappa^{\#B}$ with entries giving conditional probabilities of states of variables in A, B conditioned on states/components at the a_i . By choosing positive root distributions at the nodes a_i , and positive π_i , we ensure R will have positive diagonal entries, and hence have full rank. Furthermore, the rows of N_A, N_B are formed from the tensor product of corresponding rows of the Markov matrices on the edges of the star trees, and are thus generalized Vandermonde matrices. (Recall that if f_1, \dots, f_t are a linearly

independent set of polynomials, and u_1, \dots, u_s are points, the generalized Vandermonde matrix is the matrix $s \times t$ matrix with i, j entry $f_j(u_i)$. Here the polynomials f_j are determined by the formulae for the entries in the tensor product of the rows, and the u_i by the entries in the Markov matrices.) A generalized Vandermonde matrix has full rank for generic choices of u_1, \dots, u_s . Since $(r-1)\kappa < \min(\kappa^{\#A}, \kappa^{\#B})$, for generic parameters M_2 has rank $(r-1)\kappa$.

On the other hand, consider P_1 , where we choose all matrices on pendant edges of T_1 to be I_κ , and both the root distribution at c and M_e to have all positive entries. Then

$$M_1 := \text{Flat}_{A|B}(P_1) = N_{1,A}^T R_1 N_{1,B},$$

where R_1 is a $\kappa^2 \times \kappa^2$ diagonal matrix with entries giving the joint distribution at c and d weighted by π_1 , and $N_{1,A}, N_{1,B}$ have all zero entries except for a single 1 in each row, and full row rank. Thus M_1 has rank κ^2 . Moreover, it has at most one non-zero entry in each row and column, so both $\text{im}(M_1)$ and $\ker(M_1)$ are coordinate subspaces.

Since $\text{Flat}_{A|B}(P) = M_1 + M_2$, our goal is to show that $\text{rank}(M_1 + M_2) > r\kappa$ for generic choices of the parameters not yet specified (the Markov matrices on the trees T_2, \dots, T_r). Without loss of generality assume that $\#A \geq \#B$, so to do this it is enough to make

$$(1) \quad \text{rank}(M_1 + M_2) = \min((r-1)\kappa + \kappa^2, \kappa^{\#B}).$$

We use the following facts about matrices: Let X and Y be $s \times t$ matrices. With $\text{im}(X), \ker(X)$ denoting the image and kernel of X as a linear transformation from \mathbb{R}^t to \mathbb{R}^s , then $\text{im}(X) \cap \text{im}(Y) = 0$ implies $\ker(X+Y) = \ker X \cap \ker Y$. Also, if $\text{nullity}(X+Y) = \text{nullity}(X) + \text{nullity}(Y) - t$, then by the rank/nullity theorem $\text{rank}(X+Y) = \text{rank}(X) + \text{rank}(Y)$.

First consider the case where $(r-1)\kappa + \kappa^2 \leq \kappa^{\#B}$. By the preceding paragraph, to show equation (1) it suffices to choose parameters so that $\text{im}(M_1) \cap \text{im}(M_2) = 0$ and $\dim(\ker(M_1) \cap \ker(M_2)) = \text{nullity}(M_1) + \text{nullity}(M_2) - \kappa^{\#B}$.

Since generically N_A and N_B have full rank, it follows that $\text{im}(M_2) = \text{im}(N_A^T)$ and $\ker(M_2) = \ker(N_B)$. But $\text{im}(M_1)$ is a coordinate subspace, so it intersects $\text{im}(N_A^T)$ non-trivially if and only if the submatrix of N_A^T obtained by deleting rows corresponding to those coordinates has nontrivial kernel. That submatrix is a $(\kappa^{\#A} - \kappa^2) \times (r-1)\kappa$ generalized Vandermonde matrix with $\kappa^{\#A} - \kappa^2 \geq \kappa^{\#B} - \kappa^2 \geq (r-1)\kappa$, so it has full column rank. This proves that $\text{im}(M_1) \cap \text{im}(M_2) = 0$ generically.

Since $\ker(M_1)$ is also a coordinate subspace, its intersection with $\ker(M_2) = \ker(N_B)$ is isomorphic to the kernel of the submatrix of N_B obtained by deleting the columns corresponding to required zero entries in vectors in $\ker(M_1)$. Since this submatrix is a $(r-1)\kappa \times (\kappa^{\#B} - \kappa^2)$ generalized Vandermonde matrix, the dimension of this kernel is

$$\kappa^{\#B} - \kappa^2 - (r-1)\kappa = (\kappa^{\#B} - \kappa^2) + (\kappa^{\#B} - (r-1)\kappa) - \kappa^{\#B}.$$

Thus $\dim(\ker(M_1) \cap \ker(M_2)) = \text{nullity}(M_1) + \text{nullity}(M_2) - \kappa^{\#B}$, so $\text{rank}(M_1 + M_2) = (r-1)\kappa + \kappa^2$.

In the case where $(r-1)\kappa + \kappa^2 > \kappa^{\#B}$ the same arguments as above apply after modifying our choices so all but $\kappa^{\#B} - (r-1)\kappa$ of the entries of R_1 are zero. Then we deduce that we can choose M_2 so that $\text{rank}(M_1 + M_2) = (r-1)\kappa + \kappa^{\#B} - (r-1)\kappa = \kappa^{\#B}$. \square

Picking any internal vertex of a binary tree, the induced tripartition of the leaf variables allows us to create 3 agglomerate variables. In this way, we can view a phylogenetic model as one to which we can apply Kruskal's theorem. More specifically, consider a probability distribution P in the $GM(\kappa)$ mixture model on trees $\mathbf{T} = (T_1, \dots, T_r)$, where the T_i share a common tripartition $A|B|C$ of the leaves, arising from the vertices v_i . Suppose P_i is the weighted mixture component from T_i in P . Then from the parameters on T_i , one can give $\kappa \times \kappa^{\#A}$, $\kappa \times \kappa^{\#B}$, $\kappa \times \kappa^{\#C}$ stochastic matrices $M_{i,A}$, $M_{i,B}$, $M_{i,C}$ of conditional probabilities of states at the leaves in A , B , C , given the state at v_i . Letting $\widetilde{M}_{i,A}$ be the matrix obtained from $M_{i,A}$ by multiplying rows by the corresponding entry of the root distribution at v_i and by the weight π_i , one checks that

$$\text{Flat}_{A|B|C}(P_i) = [\widetilde{M}_{i,A}, M_{i,B}, M_{i,C}].$$

Let M_A denote the $r\kappa \times \kappa^{\#A}$ matrix obtained by stacking the $M_{i,A}$, and M_B, M_C similarly be matrices obtained by stacking the $M_{i,B}, M_{i,C}$. Then

$$\text{Flat}_{A|B|C}(P) = [M_A, M_B, M_C].$$

To apply Kruskal's theorem to this flattening, we must first show that the technical conditions on Kruskal rank of the matrices apply, at least generically.

Lemma 4.2. *Consider an r -fold $GM(\kappa)$ mixture model on trees $\mathbf{T} = (T_1, \dots, T_r)$ with a common tripartition $A|B|C$ of the leaves. Then*

$$\text{Flat}_{A|B|C}(P) = [M_A, M_B, M_C]$$

for some matrices M_A, M_B, M_C with $r\kappa$ rows. Moreover, for generic choices of the numerical parameters these matrices all have full Kruskal rank (i.e., Kruskal row rank equal to their smaller dimension).

Proof. The first claim was established in the discussion preceding the lemma.

For the second, by similar reasoning as was used in Lemma 4.1, it is enough to show one choice of parameters gives these matrices full Kruskal rank. By choosing matrix parameters on all internal edges of every T_i to be the identity matrix, we may essentially assume every T_i is the star tree, rooted at central node v_i . Choosing positive root distributions at v_i , and positive mixing parameters π_i , it then suffices to only consider one set of leaves, say A .

Now, as in the discussion of N_A in the proof of Lemma 4.1, one sees that M_A is a generalized Vandermonde matrix. Since all its submatrices are also generalized Vandermonde matrices, it generically has full Kruskal rank. \square

The next lemma allows us to tease apart distributions which arise from mixing together slices of distributions from different trees. After we have applied Kruskal's Theorem via Lemma 4.2, it will be used to identify which rows of the matrices arise from the same mixture component of the model.

Lemma 4.3 (No Shuffling Lemma). *Let T, T_1, \dots, T_r be trees with $n \geq 3$ leaves, or $n \geq 4$ leaves if $\kappa = 2$. For $i = 1, \dots, r$, let P_i be a generic probability distributions from the $GM(\kappa)$ model on the tree T_i , scaled by positive constants π_i . For a fixed choice of*

$j \in [n]$, let $A|B = \{j\}|([n] \setminus \{j\})$ and form the flattenings $\text{Flat}_{A|B}(P_i)$. Form a new matrix from any κ rows from these flattenings (with repeats allowed), and define Q so that $\text{Flat}_{A|B}(Q)$ is this matrix. Then Q does not satisfy all the phylogenetic invariants for T unless the chosen rows come from a single P_i and T is a refinement of T_i .

Proof. Note that the multiplication by the π_i has no effect on whether the tensor satisfies non-trivial invariants, because the phylogenetic varieties for the $GM(\kappa)$ model are invariant under the action of the general linear group at any leaf [8].

Consider first the case that $n = 3$, and $\kappa \geq 3$. Suppose Q is constructed from rows which come from at least two different P_i . Without loss of generality, we assume $j = 1$, so that in the notation of Theorem 3.5, the slices $Q_{(i)}$ contain the entries of Q arising from a single row of the flattening. We will show that Q does not satisfy the invariants of that theorem.

For the time being, treat two of these slices $Q_{(1)}, Q_{(2)}$ as fixed, and the third slice $Q_{(3)}$, which we may assume does not come from the same P_i as either $Q_{(1)}$ or $Q_{(2)}$, as a variable. Generically, the matrix equation

$$(2) \quad Q_{(1)} (\text{adj } Q_{(2)}) Q_{(3)} - Q_{(3)} (\text{adj } Q_{(2)}) Q_{(1)} = 0$$

then gives nonzero, linear constraints on the entries of $Q_{(3)}$.

However, for an arbitrary matrix $Q_{(3)}$ with positive entries whose sum is less than 1, we can find a P_j that has $Q_{(3)}$ as any designated slice. This shows that there exist such slices not satisfying equation (2), and hence, by Proposition 3.2, that the generic slice does not.

When $\kappa = 2$ and $n = 3$, there are no non-trivial invariants for $GM(\kappa)$ (those of Theorem 3.5 are identically zero), hence we consider $n = 4$, and use the edge invariants of Theorem 3.4. But for any choice of 4-leaf tree, and choice of index $j \in \{1, 2\}$, we can find a P_i in the tree model so that $\text{Flat}_{A|B}(P)$ has any desired generic vector as its j th row. Now Q is built from two such rows. If the P_1 and P_2 that we take these slices from are not the same, then generically, we can choose those slices to be arbitrary vectors. But then the flattening of Q with respect to the split of T will generically be a rank 4 matrix, and hence Q will not satisfy the invariants for tree T .

For larger n , the result follows from the cases above by marginalization to 3- or 4-leaf trees. \square

First we prove a theorem on the generic identifiability of numerical parameters in trees with a known common tripartition.

Theorem 4.4. *Suppose the trees $\mathbf{T} = (T_1, \dots, T_r)$ have a known common tripartition $A|B|C$, with $\#A \geq \#B \geq \#C$, and $r \leq \kappa^{\#B-1}$. If $\kappa = 2$ also suppose $\#A \geq 3$. Then both \mathbf{T} and the numerical parameters of the $GM(\kappa)$ mixture model on \mathbf{T} are generically identifiable.*

Proof. Since the trees in \mathbf{T} share a common tripartition $A|B|C$, by Lemma 4.2 if a distribution P arises from generic parameters of the model then

$$\text{Flat}_{A|B|C}(P) = [M_A, M_B, M_C],$$

where M_A , M_B , and M_C all have full Kruskal row rank, which will be $\min(r\kappa, \kappa^{\#A})$, $\min(r\kappa, \kappa^{\#B})$, and $\min(r\kappa, \kappa^{\#C})$, respectively. According to Theorem 3.1, these matrices are uniquely determined up to simultaneous permutation and scaling of the rows provided

$$(3) \quad \min(r\kappa, \kappa^{\#A}) + \min(r\kappa, \kappa^{\#B}) + \min(r\kappa, \kappa^{\#C}) \geq 2r\kappa + 2.$$

Since $\kappa \geq 2$ and $\#C \geq 1$, this inequality holds for all $r \leq \kappa^{\#B-1}$.

At this point, we have recovered the matrices M_A , M_B , and M_C up to scaling and permuting the rows. Each of the rows of the recovered M_A will have entries from a scaled slice from a tree distribution on a subtree of one of the T_i (the subtree spanning the vertex v_i and all the leaves A). We need to group these rows by the mixture components they come from. However, the No Shuffling Lemma 4.3, says that generically it is possible to do this. Since ordering the rows of M_A determines an order of the rows of M_B, M_C , we can then reassemble the flattened mixture components P_i as the product $[M_{i,A}, M_{i,B}, M_{i,C}]$ of appropriate submatrices $M_{i,A}, M_{i,B}, M_{i,C}$ of M_A, M_B, M_C .

From P_i , we recover the mixing weight π_i via

$$\pi_i = \sum_{(j_1, \dots, j_n) \in [\kappa]^n} P_i(j_1, \dots, j_n).$$

Then, by Theorem 2.1, the tree T_i and the numerical parameters on it can be identified from P_i/π_i □

Now we proceed to prove identifiability of the numerical parameters and tree parameters in our most general class of r -tree mixture models, the j -deep class.

Definition 4.5. For a positive integer j , the j -deep class of r -tuples of trees \mathbf{T} consists of all r -tuples of binary trees such that there exists a tripartition $A|B|C$ with $\#A, \#B \geq j$, $\#C \geq 1$, such that the splits $A|B \cup C$ and $A \cup C|B$ are compatible with all trees in \mathbf{T} .

Note that this definition does not require that $C|A \cup B$ be compatible with any of the trees in \mathbf{T} , so the full tripartition need not be associated to vertices in the T_i . The trees must only share two splits, each sufficiently deep in the tree. Furthermore, if \mathbf{T} is in the j -deep class, we do not assume the tripartition is known, only that it exists.

We now prove our main theorems on identifiability of parameters in r -tree mixtures. We state two versions, one for when a j -deep tripartition is known (including the case of when all the trees are known), and one for when it is not. The second of these requires a slightly stronger hypothesis on the number of mixture components.

Theorem 4.6. Suppose \mathbf{T} is in the j -deep class via a known tripartition $A|B|C$. Then both \mathbf{T} and the numerical parameters of the $GM(\kappa)$ mixture model associated to \mathbf{T} are generically identifiable provided $r \leq \kappa^{j-1}$ and either $\kappa > 2$, or $\kappa = 2$ and $\#A \geq 3$.

Proof. Fix some $c \in C$, let $D(c) = A \cup B \cup \{c\}$, and let $P_c = P|_{D(c)}$ be the marginalization of P to the leaves in $D(c)$. This is a probability tensor for the mixture of induced trees $\mathbf{T}|_{D(c)}$, with numerical parameters obtained by restricting to these induced trees. Note that the trees in $\mathbf{T}|_{D(c)}$ share the common tripartition $A|B|\{c\}$. Thus Theorem 4.4 applies

to identify the trees $\mathbf{T}|_{D(c)}$ and numerical parameters on them. Then by Lemma 4.2 we may write

$$\text{Flat}_{A|B|\{c\}}(P_c) = [M_A, M_B, M_c],$$

and for generic choices of the numerical parameters, these matrices all have full Kruskal row rank. We may further specify that the rows of these matrices, in particular M_A , have been ordered into r blocks of κ rows, corresponding to the various mixture components.

Note that since the matrix M_A has full Kruskal row rank and is $r\kappa \times \kappa^{\#A}$ with $r\kappa \leq \kappa^{\#A}$, it has full row rank. Thus we may compute a left inverse Q_A , with $M_A Q_A = I_{r\kappa}$, the $r\kappa \times r\kappa$ identity.

Returning to the consideration of the full distribution P and trees \mathbf{T} , we use Q_A to disentangle the mixture components. In each T_i let w_i be the node in the subtree spanning A through which this subtree is connected to all other leaves. Then

$$\text{Flat}_{B \cup C|A}(P) = M_{B \cup C}^T \Pi \widetilde{M}_A,$$

where \widetilde{M}_A , $M_{B \cup C}$ are stochastic matrices of probabilities of states at the leaves in A , $B \cup C$ conditioned on components and states at the w_i , and Π is a diagonal matrix with entries the product of the mixing weights, π_i , and the root distributions at w_i . While the ordering of the mixture components and root states in these matrices is arbitrary, we may assume it is the same as in the rows of M_A . Then

$$M_A = R \widetilde{M}_A,$$

where R is a block diagonal matrix whose i th block gives conditional probabilities of state changes from v_i to w_i on T_i , and is generically invertible.

Thus

$$\text{Flat}_{B \cup C|A}(P) Q_A = M_{B \cup C}^T \Pi R^{-1} M_A Q_A = M_{B \cup C}^T \Pi R^{-1}.$$

This shows that by taking the columns of $\text{Flat}_{B \cup C|A}(P) Q_A$ in blocks of κ we obtain entries associated to only one mixture component at a time. Moreover, multiplying a block of these columns by the corresponding block of rows of $M_A = R \widetilde{M}_A$, we obtain a flattened form of a single mixture component $\pi_i P_i$.

Summing the entries of $\pi_i P_i$ identifies π_i , and hence P_i . Then by Theorem 2.1 the tree T_i and the numerical parameters on it are identifiable. \square

Theorem 4.7. *Suppose \mathbf{T} is in the j -deep class. Then both \mathbf{T} and the numerical parameters of the $GM(\kappa)$ mixture model associated to \mathbf{T} are generically identifiable provided $r < \kappa^{j-1}$.*

Proof. Since the \mathbf{T} is in the j -deep class and $r\kappa < \kappa^{\#A}, \kappa^{\#B}$, for generic parameters we can use the edge invariants of Lemma 4.1 to find two splits $A|B \cup C$ and $B|A \cup C$ compatible with all trees in \mathbf{T} , with $\#A \geq \#B \geq j$, $\#C \geq 1$, simply by testing for all splits of an appropriate size.

If $\kappa = 2$, then $2 \leq r < \kappa^{j-1}$ implies $j \geq 3$, so $\#A \geq 3$. Thus for any $\kappa \geq 2$, Theorem 4.6 applies to give the conclusion. \square

We are now in a position to deduce Theorem 1.1, which will follow from Theorem 4.7 and the following lemma.

Lemma 4.8. *Let T be an unrooted binary tree with $n \geq 3$ leaves. Then there exists an internal vertex v in T inducing a tripartition $A|B|C$ such that two of the three components contain at least $\lceil n/4 \rceil$ leaves of T .*

Proof. According to Exercise 1.5 in [19], every tree has a centroid v , which is an internal node such that each component of $T \setminus v$ has at most $|V|/2$ vertices where V is the number of vertices of T . This same statement holds if we replace V with n and vertices with leaves in the definition of the centroid. Since the tree T is binary and v is an internal vertex, there are three components of $T \setminus v$. The largest component has at least $\lceil n/3 \rceil$ leaves and at most $\lfloor n/2 \rfloor$. Thus there are at least $\lceil n/2 \rceil$ leaves remaining between the other two components, which implies that in the most balanced case, one of the other two components has at least $\lceil n/4 \rceil$ leaves. Since $\lceil n/3 \rceil \geq \lceil n/4 \rceil$ this proves the claim. \square

Simple examples show the bound $\lceil n/4 \rceil$ in this lemma is the best possible.

Proof of Theorem 1.1. According to Lemma 4.8, there is an internal vertex of T inducing a tripartition $A|B|C$ such that $\#A \geq \#B \geq \lceil n/4 \rceil$ and $\#C \geq 1$. Thus $\mathbf{T} = (T, \dots, T)$ is in the $\lceil n/4 \rceil$ -deep class. Theorem 4.7 then applies. \square

5. FURTHER DIRECTIONS

The techniques employed in this paper have been primarily concerned with, and are effective for, the identification of parameters in mixture models where the underlying trees share large common substructures. Establishing identifiability of either numerical or tree parameters in situations where there is no commonality between the trees remains an open problem.

Even in the case of general Markov mixtures of two 4-leaf trees little is understood: First, in the case of two different tree topologies being mixed, it is unknown if the tree parameters are generically identifiable. Second, if the two trees are given, it is unknown if numerical parameters are generically identifiable. These problems might be addressed by finding stronger versions of the tensor rank results we have employed (*e.g.*, a strengthened version of Kruskal's theorem). But it also seems likely that a solution to these these problems will require the development of new mathematical techniques.

ACKNOWLEDGEMENT

Thanks to John Huelsenbeck for stimulating this work through describing his own investigations with mixture models with many components.

John Rhodes was partially supported by US National Science Foundation (DMS 0714830). Seth Sullivant was partially supported by the David and Lucille Packard Foundation and the US National Science Foundation (DMS 0954865).

REFERENCES

- [1] E. S. Allman, C. Ané, and J. A. Rhodes. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv. in Appl. Probab.*, 40:229–249, 2008. [arXiv:0709.0531](#).
- [2] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 2009.

- [3] E. S. Allman, C. Matias, and J. A. Rhodes. Parameter identifiability in a class of random graph mixture models, 2010.
- [4] E. S. Allman, S. Petrovic, J. A. Rhodes, and S. Sullivant. Identifiability of two-tree mixtures for group-based models. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2010.
- [5] E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186(2):113–144, 2003.
- [6] E. S. Allman and J. A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13(5):1101–1113, 2006.
- [7] E. S. Allman and J. A. Rhodes. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites. *Math. Biosci.*, 211(1):18–33, 2008.
- [8] E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.*, 40(2), 2008.
- [9] E. S. Allman and J. A. Rhodes. The identifiability of covarion models in phylogenetics. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6(1):76–88, 2009.
- [10] J. Chai and E. A. Housworth. On Rogers’s Proof of Identifiability for the GTR + Gamma + I Model, 2010. Preprint.
- [11] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.
- [12] D. Cox, J. Little, and D. O’Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, second edition, 1997.
- [13] J. H. Degnan and L. A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.
- [14] J. Felsenstein. *Inferring Phylogenies*. Sinauer and Associates, 2004.
- [15] J. P. Huelsenbeck and M. A. Suchard. A nonparametric method for accommodating and testing across-site rate variation. *Syst. Biol.*, 56(6):975–987, 2007.
- [16] J. B. Kruskal. More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293, 1976.
- [17] J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.*, 18(2):95–138, 1977.
- [18] M. Pagel and A. Meade. Mixture models in phylogenetic inference. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 121–142. Oxford University Press, Oxford, 2005.
- [19] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.
- [20] V. Strassen. Rank and optimal computation of generic tensors. *Linear Algebra Appl.*, 52/53:645–685, 1983.
- [21] D. Štefankovič and E. Vigoda. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.*, 14(2):156–189, 2007.
- [22] J. Wakeley. *Coalescent Theory*. Roberts and Company, 2008.

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF ALASKA, FAIRBANKS AK 99775
E-mail address: j.rhodes@alaska.edu

DEPARTMENT OF MATHEMATICS, NORTH CAROLINA STATE UNIVERSITY, RALEIGH, NC 27695
E-mail address: smsulli2@ncsu.edu